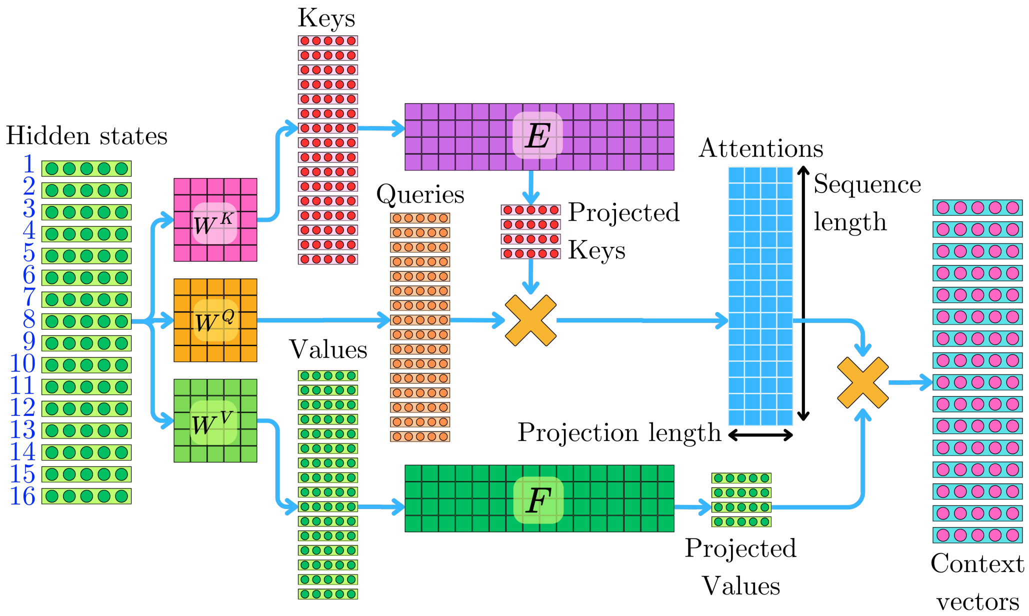
# Low-Rank Projection of Attention Matrices: Linformer

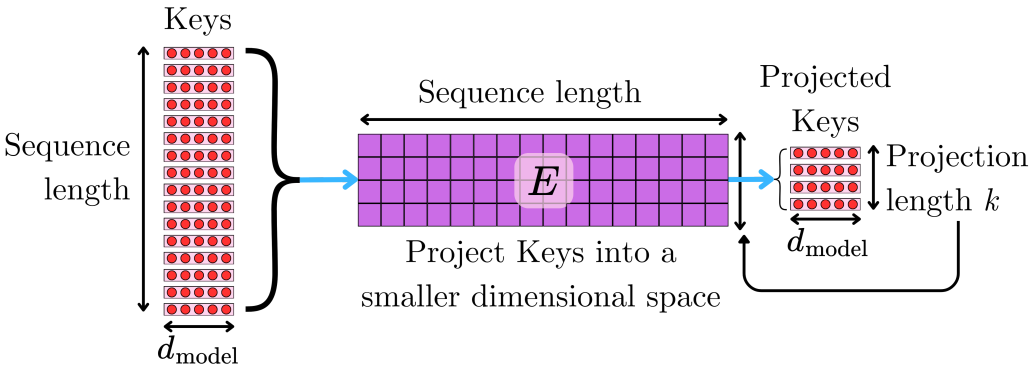
Damien Benveniste




Linear attention mechanisms represent a paradigm shift in transformer architecture by mathematically re-engineering the attention operation to achieve $\mathcal{O}(n)$ complexity while maintaining global context awareness. Linformer introduced the idea that the token-token interaction matrix could be compressed into a smaller representation without too much information loss. Instead of computing the full $N \times N$ interaction $\frac{Q^\top K}{\sqrt{d_{\text{model}}}}$ (ignoring heads for simplicity), we could first project $K$ into a lower rank dimension $k$, and compute the lower rank $N \times k$ approximation:

$$\frac{Q^\top E K}{\sqrt{d_{\text{model}}}} \tag{1}$$

where $E$ is a $N \times k$ projection matrix that project $K$ from the original dimension $d_{\text{model}} \times N$ to $d_{\text{model}} \times k$. This leads to $N \times k$ alignment score and attention matrices.




When we project with $E$, the approximation leads to the error:

$$\text{error} = \left| \frac{Q^\top K}{\sqrt{d_{\text{model}}}} - \frac{Q^\top EK}{\sqrt{d_{\text{model}}}} \right| \tag{2}$$

If the elements of $E$ follow a Gaussian distribution $\mathcal{N}(0, 1/k)$, the Johnson–Lindenstrauss lemma guarantees that:

$$P\left[\text{error} > \epsilon\right] \leq e^{-\gamma\epsilon^2 k}. \tag{3}$$

This means that the probability that we choose $E$ such that the error is greater than $\epsilon$ is bounded by $e^{-\gamma\epsilon^2 k}$. If we choose $k \to \infty$, then $P\left[\text{error} > \epsilon\right] \to 0$ for any $\epsilon$. A good choice is $k \propto \log N/\epsilon^2$, yielding:

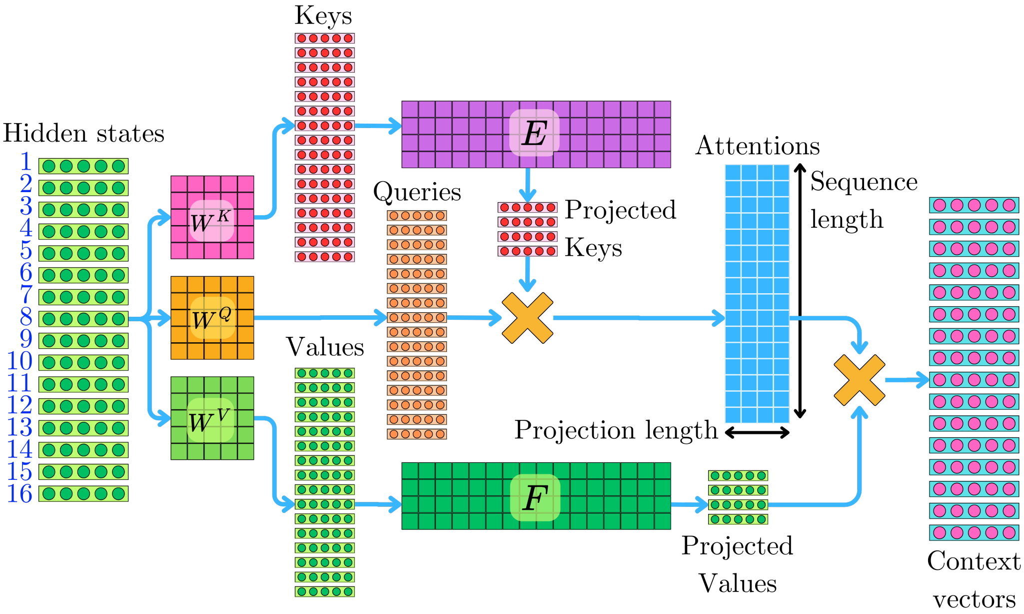
$$P\left[\text{error} > \epsilon\right] \leq N^{-\gamma}. \tag{4}$$

This means that we can choose an arbitrarily small $\epsilon$ such that $P\left[\text{error} > \epsilon\right] \to 0$ as the sequence length increases $N \to \infty$. Understand this as a mere theoretical guide that tells us that choosing $k \propto \log N$ will guarantee smaller errors as $N$ increases. In practice, $k$ is chosen independently of $N$, leading to the $\mathcal{O}(N)$ linear complexity while accepting the cost of the approximation error. Additionally, $E$ is chosen as a parameter layer for the model to learn. For example, they showed that choosing $k = 64$ with $N = 512$ leads to slightly worse performance than the full attention.

Since the attention matrix has dimension $N \times k$, we also need to project the values:

$$C = \text{Softmax}\left(\frac{Q^\top EK}{\sqrt{d_{\text{model}}}}\right) FV \tag{5}$$

where $F$ is the $N \times k$ projection matrix for the tensor $V$. As for $E$, $F$ is also learned during training.




Projecting the keys and values $EK$, $FV$ leads to complexity $\mathcal{O}(Nk)$. Computing the alignment scores $Q^\top EK$ and the context vectors $C = AFV$ are also following $\mathcal{O}(Nk)$. Since we fix $k$, the overall time and space complexity is $\mathcal{O}(N)$